基于多源数据的领域主题演化路径分析

张敬[1,2]，朱相丽[1,2*]

（1.中国科学院文献情报中心，北京 100190；2.中国科学院大学经济与管理学院信息资源管理系，北京 100190)）

摘要：[目的/意义] 为了全面、客观、高效、直观地掌握科技领域主题的发展规律和演变趋势，本文提出了一种基于多源数据的领域主题演化路径识别和分析框架。**[方法/过程]**获取不同来源的科技文献数据，利用多维样本有序聚类方法辅助时间切片，基于改进的词袋构建方法，提升 LDA 模型主题识别效果，借助 Louvain 社区发现算法在主题层进行多源数据的融合，分析领域主题演化路径。**[结果/结论]**利用美国太赫兹研究领域基金项目、论文和专利三种来源的数据进行实证研究，结果表明，三种数据源能够清晰划分出 4 个时间窗口，改进的词袋构建方法能够表征更准确的领域信息内涵，主题社区有助于从多源数据复杂的演化网络中厘清主题演化脉络。

关键词：多源数据融合；领域主题演化路径；LDA 主题模型；词袋构建；时间窗口划分；有序样本聚类；Louvain 社区发现

分类号：G301

Analysis of domain topic evolution path based on multi-source data

Zhang jing[1,2], ZHU Xiangli[1,2 *]

（1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China； 2. Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China）

Abstract: [Purpose/significance] In order to comprehensively, objectively, efficiently and intuitively grasp the development law and evolution trend of domain topics, this paper proposes a framework for identifying and analyzing the evolution path of domain topics based on multi-source data. **[Method/Process]** Acquire scientific and technological literature data from different sources, use multi-dimensional ordered sample clustering method to assist temporal slicing, enhance the LDA model topic identification effect based on an improved word packet construction method, utilize Louvain Community Detection Algorithm for fusion of multi-source data at the topic level, and analyze domain topic evolution path. **[Results/Conclusion]** The empirical study in terahertz research field in

the United States was conducted using the data from three sources about Fund project, the paper and the patent. The results show different sources are clearly divided into four unique development stages, and the improved word-bag construction method could represent more accurate domain information, and simplified topic communities can help extract evolution paths from complex networks.

Keywords: multi-source data fusion; domain topic evolution path; LDA model; Word bag construction; time window division; multi-dimensional ordered sample clustering; Louvain Community Detection Algorithm

1 引言

世界科技发展态势深刻变化，国际科技创新环境与竞争格局加速调整。科学技术发展呈现动态性，学科领域的研究主题不断演进。综合利用多源数据信息，动态跟踪学科领域主题演进，能够有效揭示学科领域知识发展变化及其相互作用的特征和规律，从而追溯学科发展轨迹、发现新的知识增长点，进而以超前的思维和战略决策引导科技领域的发展。基于多源数据挖掘和认识领域主题发展规律与演化趋势，不仅对科研人员全面把握领域发展脉络、发展现状和未来趋势具有基础性作用，还能够为政策制定者预测科技前沿、部署创新战略提供重要情报保障。

本文面向科技领域主题演化分析工作，提出基于多源数据的领域主题演化分析框架，利用基金项目、论文和专利信息，全面准确地识别领域重要研究主题，分析领域发展态势，把握不同研究主题的演化趋势。

2 相关研究

2.1 多源数据研究

“多源数据”是指从不同数据来源中获得的不同类型的数据，这些数据还可能具有不同的实体类型[1]，其多源性体现在“数据来源类型-数据类型-实体类型”三个层次中。李广建和杨林[2]指出，同一个事实或规律可以同时隐藏在不同的数据形式中，也可能是每一种数据形式分别支持了同一个事实或规律的某一个或几个侧面，这既为数据和信息分析的结论的交叉验证提供了契机，也要求分析者在分析研究过程中有意识地融集各种类型的数据。

多源数据的有效利用必须进行科学合理的数据融合。H. Y. Xu 等[3]提出科学计量学领域多源数据融合的三个过程，即前期的数据类型融合，中期的数据关系融合和后期的集合聚类。谭晓和李辉[4] 将多实体和多关系融合应用到主题关联，并利用图模型识别社区结构，构建多源数据知识融合框架。冯佳等[5] 提出面向研究前沿识别的载体-特征-关系融合模型，用于实现基于多源数据和深入语义层面的研究前沿识别。X. Wang [6] 采用相关分析、综合因子分析、熵权法、理想解

相似性顺序偏好技术（TOPSIS）法和二维四象限映射法，从多维度、多因素、多指标、多方法融合的角度对期刊的话语权进行评价。陈启明等[7] 分别识别新闻主题和政策主题，拟合时间和主题相似度因素，探索突发公共事件主题的政策趋向规律。胡吉霞[8] 利用聚类算法和图卷积自编码网络模型进行网络节点对齐和网络结构融合，生成融合主题、关键词和实体的多维度学科知识网络，揭示学科知识的静态结构。

研究发现，当前对于各种数据源的独特性质以及不同数据源之间的关系认识还不够明确，难以解决不同数据源术语表达和语义不一致的问题，在实际应用中多元关系融合较为复杂，多源数据融合尚未得到广泛应用。因此，基于多源数据的主题识别和主题演化分析需要利用不同数据源的特征，借助多种手段降低不同数据源的信息差异所带来的负面影响，探索更加简洁高效的多源数据融合方案。

2.2 主题演化研究

主题演化是“以词语为表征的学科主题在时间维度上的发展变化过程”，体现了研究主题的新陈代谢规律，蕴含着学科领域的发展态势和未来走向[9] 。主题演化分析过程一般包括数据获取、时间窗口划分、主题识别、主题关联和主题演化分析五个步骤。

国内外学者对主题演化分析的研究维度，主要分为主题强度的演化和主题内容的演化两个方面[10] ，现有研究多数采用可视化方法构建和分析主题演化路径。通过可视化的方式展现主题演化网，能够生动形象地揭示演化脉络，增强研究人员的洞察力和感知力，是有效分析海量信息的重要途径。2003 年，美国国家研究院提出科学知识图谱的概念[11] ，S. Morris 等 [12] 率先以时间线图谱的方式分析和展现研究前沿主题的演化情况。G. Palla 等 [13] 提出社区网络演化过程中可能存在的演化形式，包括新生、消亡、扩张、收缩、融合和分裂六种。近年来，通过多种方法简化主题演化网络、利用桑基图和河流图等方式呈现主题演化路径的可视化方案逐渐发展。周源等 [14] 提出基于主题变迁的领域发展路径识别方法，引入学者信息，利用 Kmeans++ 算法获取不同时间片上的主题，利用谱聚类的方式合并类似的主题，分析领域技术发展规律，快速定位领域发展热点和重要学者，实现领域发展河流图的全自动输出。陈悦等 [15] 通过技术群相似度时序分析法，利用桑基图展现技术融合与扩散演化路径。刘怀兰等 [16] 基于时序主题建模、时序主题关联以及河流图可视化，完成领域多源数据的融合主题挖掘和多维度技术演化路径分析。

研究发现，现有关于学科领域创新演化路径的研究，在数据获取时多选择以论文表征科学创新成果、以专利表征技术创新成果[17] ，仅从单源或二源数据中分析领域创新演化过程。时间划分方案对主题识别工作具有重要影响，但多数研究采用固定时间窗口或滑动时间窗口方法，受到主观因素和研究经验的影响；而且由于不同数据源的研究内容具有相对滞后性[18] ，统一划定的时间窗口往往不能有效应对具备不同数据特征的多源数据。主题识别时，传统分词工具提取的单词无法有效表征领域内涵，构建领域词典的方法难以高效、全面地掌握领域全部知识，现有的新词发现算法英文适用性有限。主题演化分析时，难以有效应对主题节点较多、演化关系复杂的情况，同时由于同一研究内容在不同数据源中的主题词不完全一致，不能通过直接对齐进行多角度的主题演化分析，高效、准确提取和分析清晰的主题内容演化信息存在挑战。

本文针对上述问题，提出了基于多源数据的领域主题演化路径分析框架，综

合利用基金项目、论文和专利数据，通过多维有序样本聚类方法辅助划分时间窗口，利用改进的词袋构建方法对 LDA 主题建模效果进行优化，对识别出的若干主题进行时序关联，运用 Louvain 社区发现算法简化主题演化网络，融合来自不同数据源的主题节点，应对主题演化分析过程中多源数据主题难以对齐的问题，最后从主题演化形式特征和内容特征两个角度全面、准确、高效地识别和分析领域发展过程中主题演化现象。

3. 研究设计

本文所采用的研究方法整体流程（图 1）主要包括多源数据获取、文本主题挖掘、主题演化网络构建与分析 3 个步骤，领域专家智慧融入每个步骤。其中，多源数据的融合分为两步：首先对来自不同数据源的主题进行前后时间窗口的主题关联，运用相似度计算方法；而后将属于同一研究范畴的不同来源的主题融合在同一社区，运用 Louvain 社区发现算法。

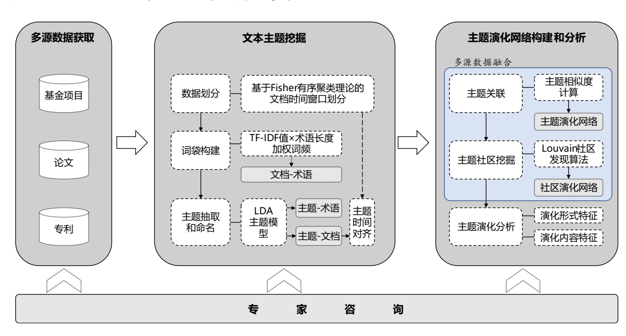


图 1 研究方法流程

3. 1 多源数据获取

多源数据的集成对领域发展态势的全面准确感知具有重要作用。在科技创新领域，基金项目数据中蕴含着领域专家和决策者共同认可的重要研究主题，期刊论文是领域基础研究成果的重要传播阵地，专利则是面向产业应用的重要技术成果载体。综合应用三种数据源能够从不同侧面发现领域重要研究内容，实现尽可能全面、准确的分析。因此，本文选取 Web of Science 数据库的 SCI 期刊论文、DII 专利文献以及 Digital Sciences 咨询公司的 Dimensions 平台（https://www.dimensions.ai/）的基金项目信息作为数据来源，主要利用标题和摘要文本进行主题挖掘，分别以论文出版时间、专利申请时间、项目开始时间作为三种数据源的时间信息。根据研究目标，选取所需字段组建初始数据集。

3. 2 文本主题挖掘

从不同来源的自然语言文本中挖掘主题进行演化分析，首先对数据集分别进行时间维度的划分，构建若干带有时间信息的语料库，通过分词手段从文本中提取具有代表性的术语，利用主题模型进行主题抽取，并将不同来源的主题进行时间对齐。

3.2.1 时间窗口划分

为解决数据集时间划分主观随意的问题，有研究者[19-20] 基于主题分布特征，对专利数据、论文数据和网页数据进行时间段划分。该方法通过主题建模，将“时间-文档”矩阵转为“时间-主题”矩阵，用多维主题特征将时间表示为向量，再通过降维和可视化的方法，将连续的时间划分成若干簇，每个簇内主题分布较为

接近，簇间则存在较大差异。考虑到不同数据源研究内容存在相对滞后性，本研究借鉴上述研究中将时间看作多维向量进行聚类的思想，利用多源数据独特的形式特征，分别进行多维有序样本聚类，最后对各时间段产生的不同主题的时间进行对齐，咨询专家意见迭代产生最终用于领域主题演化分析的时间划分方案。

对于时间序列数据而言，事物发展阶段的划分不能打乱样本时间的序列关系，只有相邻的样本才能聚到一类。1958 年 W. D. Fisher 提出用于解决此类问题的有序聚类算法[21]，其基本思想是定义类的直径，在分类必须相邻的限制条件下定义了损失函数，在逐步递推的计算中寻找到使得损失函数最小的最优分类。目前，该方法被应用于土壤学[22]、植物学[23]、地质学[24]等多个学科领域，2016 年祖坤琳等[25] 通过构建专利特征向量，基于 Fisher 有序聚类方法对专利知识的发展阶段进行划分，表现专利研究主题在不同时期的发展变化。

本文对于多源时间序列数据进行有序聚类的实验方案参考严广松和路允芳[26] 提出的多维有序样本的聚类方法，利用综合指标法将多维观测值压缩到一维空间后进行有序样本聚类，通过损失函数评估输出最佳聚类方案，实现对多个文档集的时间划分，用于后续的词袋构建和主题识别。通过将不同文档集生成的若干主题进行时间维度的对齐，划分领域整体发展阶段。时间划分的过程和结果均咨询专家意见，以保证实验的可靠性。

3.2.2 词袋构建

主题抽取的效果很大程度上受到文本分词效果的影响，传统分词方法无法有效挖掘短语，切出的词汇对文档的代表性不足。有研究者[20]利用基于 TF-IDF 的循环迭代拼接法，根据关键词左右两个方向的拼接形成短语。本文借鉴其研究思想，采用 TF-IDF 值和术语长度加权词频的方法，尝试解决这一研究难题。

TF-IDF 的核心思想是，假设某个单词或短语在一篇特定的文章中出现的频率较高，同时在数据集内其他文本中出现的频率很低，那么该术语的 TF-IDF 值较高，在数据集中类别区分能力很好，对文档的代表性很强。然而在传统的 TF-IDF 算法中，短语在文档中出现的频率远低于单词，不易被选中到候选术语集合。因此，本文利用 TF-IDF 算法，设置词频上下边界，识别每篇文档中包含 1-5 个单词的术语，并根据 TF-IDF 值和术语长度进行词频加权，构建新的文档词袋。文档中术语的加权词频计算方式为：

$$F'(w,Di) = F(w,D_i) \times TF-IDF(w,D_i) \times len(w) \qquad (1)$$

其中，$F(w, D_i)$为术语 w 在文档 D_i 中的原始词频，$TF\text{-}IDF(w, D_i)$表示 w 在文档 D_i 中的 TF-IDF 值，len(w)表示术语长度（$1 \leq len(w) \leq 5$）。

3.2.3 主题抽取和命名

LDA 主题模型是一种典型的词袋模型，常用于各领域的主题识别和主题演化分析过程。LDA 模型基于三层贝叶斯网络结构，一篇文档代表若干主题构成的一个概率分布，而每一主题又代表若干词语构成的一个概率分布，形成“文档-主题-词”的三层结构。因此，LDA 的模型计算结果可以从“文档-术语”分布中得到“主题-文档”和“主题-术语”的两个概率分布，这为后续分析过程提供丰富的数据信息。LDA 主题模型实验过程中，需要设定主题个数 K。主题数量对主题识别结果具有重大影响，一般采用困惑度（Perplexity）作为模型评估指标[27]，选取处于困惑度曲线拐点的主题数作为 K 值。

根据“主题-术语”分布矩阵，参考领域文献和专家意见进行主题命名。根据“主题-文档”分布矩阵，可以计算主题强度和主题时间。主题强度为主题支持文档与当前时间窗口下语料库全部文档的比值，表征主题研究热度。主题时间为主

题支持文档的平均时间，表征主题的新颖程度。由于不同数据源研究内容存在相对滞后，本研究利用主题时间对多源数据主题进行对齐，开展后续的主题演化分析工作。

与常规的文本主题挖掘方法相比，本文的创新之处在于：针对多源数据研究内容相对滞后的固有现象，设计时间窗口划分方案，解决当前研究中时间划分主观随意问题，并利用主题时间进行多源数据研究内容时间维度的对齐；针对分词过程中词汇代表性不足、可理解性差的问题，综合利用术语原始词频、TF-IDF 值、术语长度构建词袋，改进 LDA 模型主题识别效果。

3.3 主题演化网络构建与分析

基于多源数据主题演化研究需要对不同数据来源的主题进行融合分析。对于分别从多源数据中识别的不同主题，首先通过前后时间窗口的主题关联，选择具有高关联性的主题对构建初步的主题演化网络；而后利用 Louvain 算法将来自不同数据源的主题节点聚合成若干社区，实现主题层面的多源数据融合；最后综合利用多源数据信息，从主题社区演化形式特征和内容特征两个角度综合分析主题社区演化趋势。

3.3.1 主题关联

本文将主题视为由若干主题词表征的多维向量，利用余弦相似度算法[28]，对相邻时间窗口下的各个主题进行相似度计算。利用箱型图对计算得到的相似度进行异常值检测，选取相似度异常高的主题对，进行主题时序关联。以上步骤可根据高关联主题对建立前后时间窗口的演化联系，从而以主题为节点，初步构建领域发展的主题演化网络。

箱型图是一种用作显示一组数据分散情况的统计图，用于反映数据分布特征，对单变量大样本异常值的标记十分有效、方便和直观[29]。按照界定阈值的大小，异常值可分为轻度异常值和极端异常值两种。实际应用中，多采用轻度异常值作为异常值选定依据（如图 2），轻度异常值的上边界 $H=Q_3+1.5\times IQR$。Q_1 表示数据的第一四分位点，Q_3 表示数据的第三四分位点，IQR 表示 Q_3 和 Q_1 两数之差。

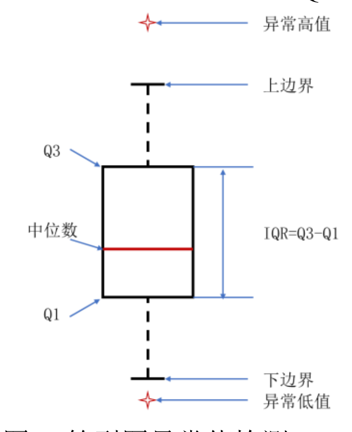


图 2 箱型图异常值检测

3.3.2 社区发现

通过主题直接构建的主题演化网络中节点多、关系复杂，同一研究内容在不同时间窗口主题的主题词不完全一致，不能通过直接对齐进行全面深入的主题演化分析。社区发现算法在复杂网络中应用较多，但在主题识别和主题演化分析工作中应用较少。在主题演化网络中通过图计算进行社区发现，一方面可以简化演化网络，清晰呈现领域发展趋势；另一方面可以融合不同来源的主题，综合利用

多源信息进行领域主题演化分析。

社区划分的算法分为分离型算法和聚合型算法，近年来也有研究者提出一些新的算法，比如基于模块度优化的算法，基于统计推理的随机游走算法，以及标签传播算法等等[30]。M. E. J. Newman [31] 首次在社区网络划分中引入了模块度（Modularity）的概念，衡量社区内节点的连边数与随机情况下的边数的差距。模块度取值范围为[-0.5, 1]，值越大表明社区结构越符合高内聚低耦合的特征，社区划分质量越高。当前模块度已成为社区划分中应用最广泛的评价函数，实际应用中模块度一般在 0.3-0.8 之间。Louvain 算法就是基于模块度划分的社区发现算法，在实现过程中包括节点移动和社区聚合两个阶段[32]，模块度可以在该算法运行的每一步衡量产生的社区是否为相对最佳的划分结果，最终评估输出最佳的社区划分方案。

因此，针对大规模主题演化网络结构复杂、理解成本高的问题，以及多源主题融合的需要，本文利用社区发现算法对初步构建的主题演化网络进一步划分，将主题节点聚合成若干社区，实现多源数据主题层次的融合，构建简洁的多源主题融合演化网络。

3.3.3 主题演化分析

多源主题融合演化网络形成了“主题词-主题-社区”的三级结构，主题社区在时间维度上形成清晰的演化路径，能够反映领域发展脉络，揭示不同研究分支的发展趋势。因此，本文从形式特征和内容特征两个角度，对主题社区演化展开分析。

（1）主题演化形式特征

G. Palla [13] 提出社区网络演化过程中可能存在新生、消亡、扩张、收缩、融合和分裂六种演化形式（如图 3），新生和消亡即网络的出现和消失，扩张和收缩即网络内部节点数量增长和减少，融合和分裂分别表征该网络与前/后一时间窗口的其他网络产生联系。在此基础上，图书情报界学者将这六种演化形式应用于研究主题演化过程中 [33-35]，在时间维度上，对主题演化进行推理分析。

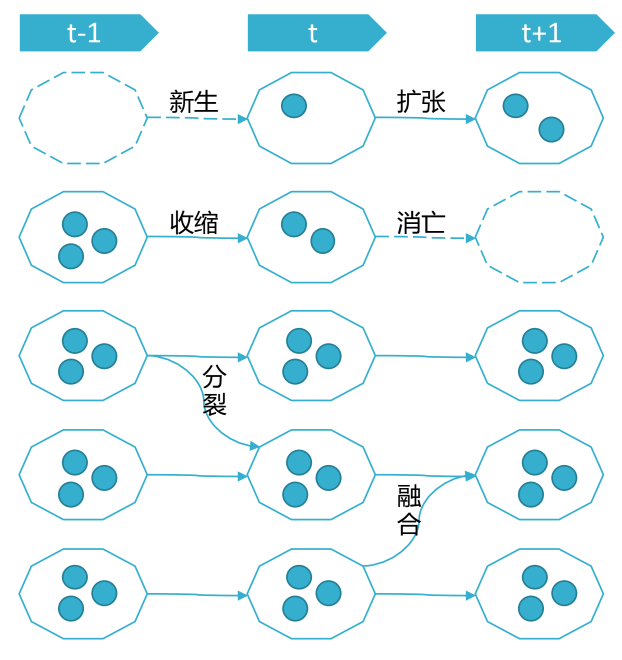


图 3 网络演化形式

主题演化形式特征分析研究不同时间窗口下社区内部和社区之间的关联关

系。主题社区内部不同时间窗口下存在若干主题节点，通过主题节点数量的变化形成新生、消亡、扩张、收缩四种演化形式。社区之间通过高关联主题对建立了前后时间窗口下社区间的演化联系，形成分裂和融合两种演化形式。在此基础上，不同数据源具有自身独特的性质，基金项目数据代表受到领域专家和决策者一致认可的研究内容，论文数据和专利数据则分别代表基础研究与应用研究的重要成果。对于社区内不同数据来源的主题节点发展趋势进行分析，能够从更广阔的视角进一步丰富研究结论。

3.3.4 战略坐标图

战略坐标图是基于研究主题或聚类，描述各研究主题的发展状况和演变趋势的方法。战略坐标图以向心度为横坐标轴，以密度为纵坐标轴，以两者的中位数或均值为坐标原点，将研究主题簇表示在平面坐标系中[36]。本文所采用的密度和向心度的计算方法参考 B. Lee 和 Y. I. Jeong 发表于 2008 年的研究论文 [37]，密度指标反映一个研究主题簇的内部聚合能力，向心度指标反应研究主题簇与其他研究主题簇的连接能力。密度越大，说明该研究主题内部结构稳定性越高；向心度越大，说明该研究主题簇在整个研究领域中处于越核心的地位。

主题演化内容分析借助战略坐标图研究不同社区发展演变趋势。具体而言，将主题社区根据时间窗口划分为不同的主题簇，在全局网络下利用各主题簇所包含的主题词分别计算密度和向心度，将其分布在战略坐标图的四个象限（图 4）。第一象限主题簇核心且稳定，是本领域研究热点和重点，受到广泛关注，内部结构稳定；第二象限主题簇为边缘、稳定类，内部结构稳定，但与其他主题簇联系松散，研究相对独立；第三象限主题簇是边缘且非稳定类，内部结构松散，研究相对不成熟，在领域中处于边缘位置；第四象限分布着核心但不稳定类主题簇，是领域研究的活跃分支，但内部结构不稳定，发展尚不成熟。随着领域的发展，各主题社区研究内容呈现不同的发展趋势，密度的提升表征自身发展趋向稳定，向心度的提升表明其在领域发展的过程中占据愈发核心的位置。

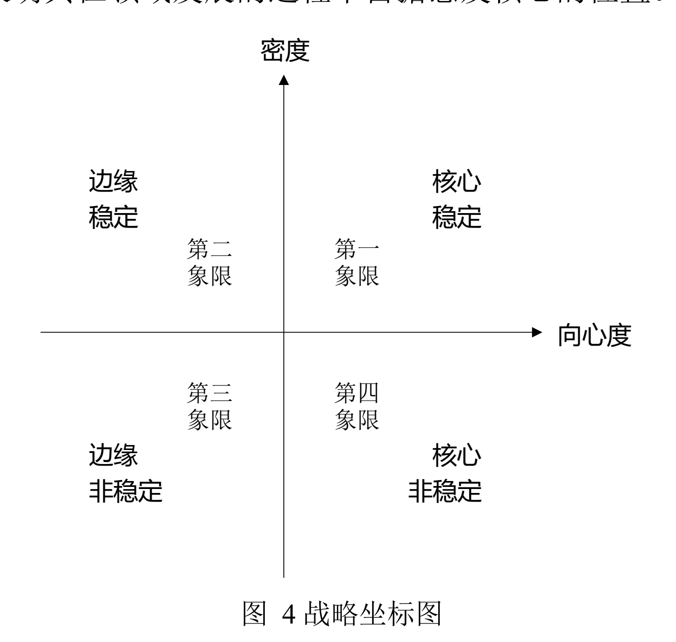


图 4 战略坐标图

4 实验与结果分析

为了验证研究框架的合理性和有效性，本文选择美国太赫兹领域进行实证研

究。太赫兹以其独特的性能和广泛的应用越来越受到各国的关注，被国际科学界公认为是高科技领域的必争之地，呈现出基础科学、先进技术和产业化三方面多元化快速发展的新局面。

4.1 多源数据获取

本文分别选取 Dimensions 平台、Web of Science 数据库的 SCI 期刊论文、DII 专利文献，以 “terahertz”为主题词，检索获取论文通讯作者或第一作者来自美国、专利申请人和发明人均来自美国、基金项目资助国和研究机构所在地均为美国的相关文档，建立初始数据集。最终，共获得 1255 份基金项目文献、5121 份期刊论文文献、1204 个简单专利同族。

4.2 领域主题识别

4.2.1 文档时间窗口划分

论文以年度发文量和年度被引频次、专利以年度申请量和年度申请人数量、基金项目以年度新增项目数量和年度项目金额为特征，分别划分时间窗口。通过对时间序列数据进行多维有序样本聚类，利用 Python 库的 Kneed 包定量选取损失函数数据曲线的拐点[38]，自动输出时间划分最佳方案，不同数据源的时间窗口划分结果如图 5 和表 1 所示。

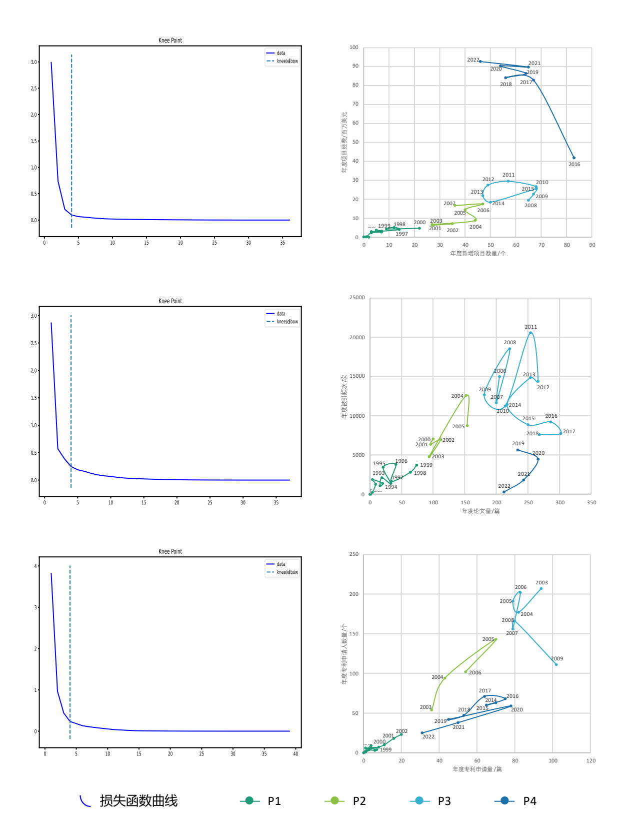


图 5 领域不同数据源文档时间聚类图

表 1 领域不同数据源的时间窗口

时间窗口	基金项目	论文	专利
P1	1987-2000	1986-1999	1984-2002
P2	2001-2007	2000-2005	2003-2006

P3	2008-2015	2006-2018	2007-2013
P4	2016-2020	2019-2022	2014-2022

图 5 上、中、下分别为基金项目、论文和专利，左为损失函数曲线，右为时间分布散点图。由图 5 和表 1 可知，基金项目、论文和专利可以分别划分为 4 个发展阶段。美国太赫兹领域最早源于 1984 年的专利文献，约 2000 年之前是该领域发展的第一阶段，第二阶段约在 2001-2006 年，第三阶段于 2007 年左右开始。2014 年之后，专利、基金项目和论文先后进入第四阶段。

4.2.2 词袋构建

对同一数据源、同一时间窗口下的文档进行文本处理，利用 3.2.2 所述方法，基于 TF-IDF 值和术语长度加权词频构建词袋，改进 LDA 模型主题识别效果。首先，构建领域停用词表和停用短语表，利用 Python 的 NLTK 库进行文档分词，将停用词和标点替换为与领域无关的特殊标记词；其次，利用 TF-IDF 算法，抽取单词或短语作为候选术语，具体参数设置为：max_df=0.8, min_df=0.01, ngram_range=(1,5)，即抽取长度范围为 1~5 的单词或短语为候选术语，并将其在文档中出现的频率限制在 1%~80%。再次，剔除包含特殊标记词的候选术语，删除停用短语表中的候选术语，形成术语集合。最后，根据术语在文档中原始出现频次、TF-IDF 值和术语长度，计算术语加权词频，构建词袋。最终构建的词袋中术语长度大于 2 的短语占比超过 76.94%。

4.2.3 主题抽取

按照构建好的词袋，设置参数，利用 LDA 主题模型进行主题抽取。本文设定 α =50/K、 β =0.01、迭代次数为 500 次，并设定每篇文档输出概率超过 0.1 的所有主题，每个主题输出概率最高的 100 个词（包括概率与第 100 个词一致的所有词汇）。根据困惑度指标，计算最佳主题数 K，不同数据源、不同时间阶段的困惑度曲线如图 6 所示（左、中和右依次为基金项目、论文和专利，由上至下为 P1-P4），其文档数和最佳主题数如表 2 所示。

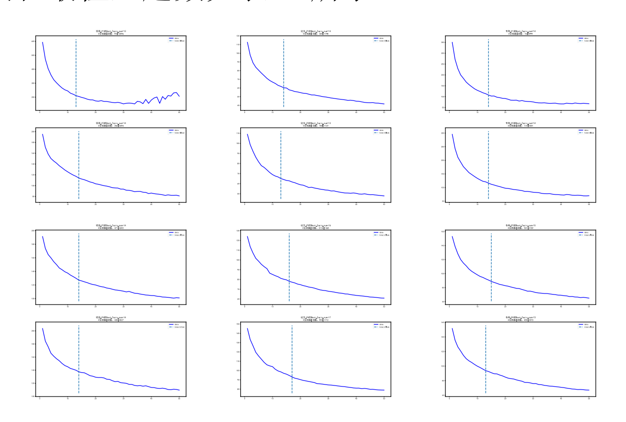


图 6 领域不同数据源不同时间窗口困惑度曲线

表 2 领域不同数据源不同时间窗口的文档数和主题数

数据来源	基金项目				论文				专利			
时间窗口	P1	P2	P3	P4	P1	P2	P3	P4	P1	P2	P3	P4
文档数量	93	256	471	435	340	708	3118	955	76	172	501	455
主题数	13	14	14	14	14	13	16	17	14	14	15	13

LDA 模型抽取的各主题命名结果（部分）如表 3 所示。通过咨询领域专家认为，美国太赫兹领域三种数据源多个时间窗口形成的 171 个主题，基本涵盖太赫兹领域的关键术语，主题命名相对高效、准确。不同数据源不同时间窗口的主题时间-主题强度分布如图 7 所示。三种数据源分别在四个时间窗口下形成不同的研究主题，根据主题时间，分别以 2002、2007 和 2015 年为分界点，数据集自然呈现 1984-2002、2002-2007、2007-2015、2015-2022 四个发展阶段。通过阅读领域大量文献以及咨询专家意见，上述阶段划分符合领域实际发展情况。

表 3 领域不同数据源不同时间窗口的主题详细信息

主题序号	时间窗口_数据源类型	主题名	部分主题词
1	P1 论文	自由空间电光采样	'electro optic', 'free space', 'free space electro optic', 'freely propagate terahertz', 'electro optic sampling', 'optic', 'terahertz pulse', 'freely propagate', 'space electro optic', 'free space electro'
2	P1 论文	太赫兹光学非对称解复用器	'optical', 'terahertz optical asymmetric demultiplexer', 'optical asymmetric demultiplexer', 'terahertz optical asymmetric', 'network', 'terahertz optical', 'optical amplifier', 'switch', 'asymmetric demultiplexer', 'optical asymmetric'
3	P1 论文	太赫兹脉冲	'terahertz pulse', 'cycle terahertz pulse', 'cycle terahertz', 'pulse', 'cycle', 'half cycle', 'optical', 'optical frequency comb', 'single cycle', 'diffraction'
……	……	……	……
169	P4 项目	太赫兹时域光谱厚度测量技术	'terahertz spectroscopy', 'time domain', 'thickness', 'measurement', 'time domain terahertz', 'spectroscopy', 'thickness measurement', 'domain terahertz', 'measurement system', 'instrument'
170	P4 项目	太赫兹成像系统	'image', 'terahertz image', 'imaging', 'high resolution', 'image system', 'terahertz image system', 'high energy', 'terahertz imaging', 'detection', 'sensor'
171	P4 项目	太赫兹无线通信网络	'network', 'wireless', 'communication', 'terahertz band', 'data rate', 'terahertz communication', 'data', 'spectrum', 'band', 'high data rate'

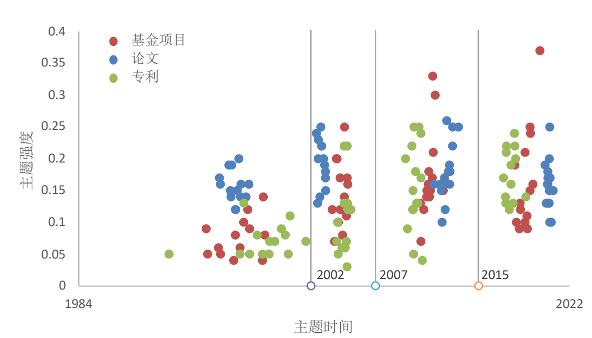


图 7 领域不同数据源不同时间窗口的主题时间-主题强度分布

4.3 主题演化网络构建与分析

4.3.1 主题演化网络构建

对相邻时间窗口下的各个主题进行相似度计算，利用 3.3.1 所述箱型图检测算法，选取相似度大于 0.177 的主题对（如图 8），进行主题时序关联，最终形成 210 对高关联主题。根据这些主题对，初步构建领域发展的主题演化网络，利用

Louvain 算法实现社区发现（模块度 Q>0.62）。为保证合理的社区规模，本文设置社区内部主题节点数量最低为 3，最终形成 8 个关系紧密的主题社区（表 4）。图 9 展示了调整后的主题演化网络图，横向维度表示时间的演进，纵向维度表示不同的主题社区，节点大小代表主题强度，边的宽度代表演化关联的紧密性。分析发现，美国太赫兹领域主要研究主题社区为：太赫兹辐射源、太赫兹探测器、太赫兹量子级联激光器、太赫兹时域光谱建模与分析、太赫兹电子学、太赫兹通信、太赫兹检测成像、太赫兹功能器件制备材料。

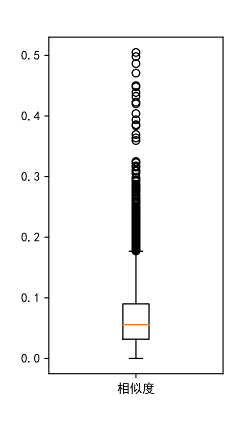


图 8 领域主题相似度箱型图

表 4 Louvain 算法初步识别领域主题社区

社区编号	主题节点	社区命名
#1	1, 3, 8, 12, 13, 44, 47, 50, 74, 93, 97, 122, 153, 160	太赫兹辐射源
#2	4, 9, 11, 14, 43, 46, 53, 77, 81, 85, 88, 123, 136	太赫兹探测器
#3	6, 42, 55, 59, 71, 82, 87, 92, 99, 100, 103, 106, 108, 109, 112, 121, 132, 139, 142, 144, 145, 150, 151, 152, 157, 161	太赫兹量子级联激光器
#4	5, 10, 45, 51, 54, 72, 84, 86, 91, 111, 117, 124, 129, 135, 137, 154	太赫兹时域光谱建模与分析
#5	7, 49, 70, 78, 80, 94, 114, 118, 119, 120, 126, 130, 158, 159	太赫兹电子学
#6	95, 105, 115, 128, 131, 147, 148, 156, 166	太赫兹通信
#7	69, 83, 101, 110, 116, 138, 146, 162, 164	太赫兹检测成像
#8	52, 57, 89, 90, 96, 98, 107, 125, 133, 134, 140, 141, 143, 155	太赫兹功能器件制备材料

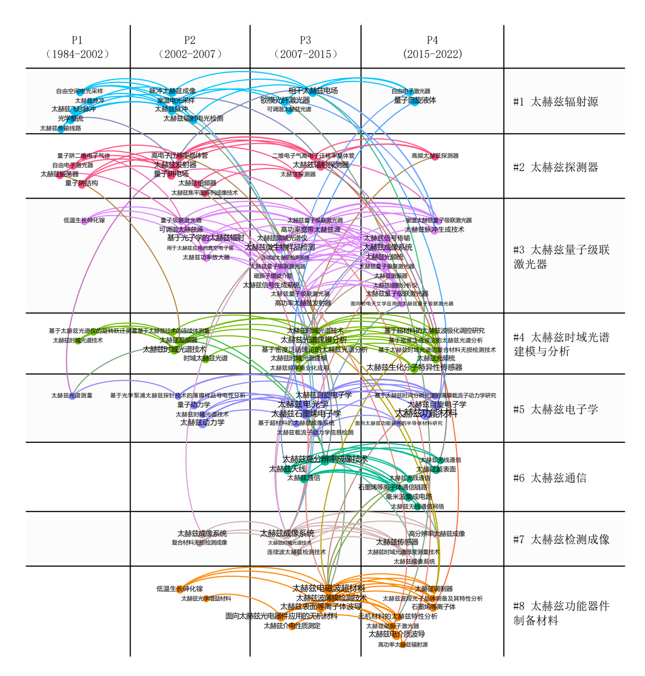


图 9 领域主题演化网络

4.3.2 主题社区演化分析

本文构建主题社区演化图，分析主题社区演化形式特征和内容特征，并在此基础上咨询专家意见，对实验结果的有效性进行评估和验证。

（1）演化形式特征分析

根据社区主题节点数量时序分布（图 10）和主题社区时序关联图（图 11）分析发现：

从社区内数据源类型（图 10）来看，多源数据在不同主题社区中的分布是不均衡的，领域研究初期论文在主题社区内外产生的演化影响较大，中后期其他数据源逐渐扩大影响力。#1 太赫兹辐射源前期以论文主题为主，并向基金项目主题或专利主题发展。#2 太赫兹探测器和#4 太赫兹时域光谱建模与分析以论文主题研究为主，基金项目主题为辅。#3 太赫兹量子级联激光器和#6 太赫兹通信主题社区内三种数据源的主题均具有重要作用，但前者专利主题节点数量显著增长，后者则呈现较为均衡发展的态势。#5 太赫兹电子学中基金项目主题总体较多，论文主题数量逐渐增长。#7 太赫兹检测成像和#8 太赫兹功能器件制备材料则分别呈现基金项目和专利主题、论文和专利主题均衡发展的趋势。

从社区内不同时间段的主题节点数量（图 10）来看，不同主题社区存在收缩、扩张和平稳发展三种状态。#1 太赫兹辐射源、#2 太赫兹探测器的研究主要集中在领域发展前期，随着时间推移主题节点数量逐渐减少，社区发展呈逐渐收缩的发展态势。#3 太赫兹量子级联激光器、#6 太赫兹通信、#7 太赫兹检测成像、#8 太赫兹功能器件制备材料主题节点数量增长明显，处于持续快速扩张阶段。#4 太赫兹时域光谱建模与分析、#5 太赫兹电子学主题社区研究起步早、主题节点数量持续增长，近期处于平稳发展阶段。

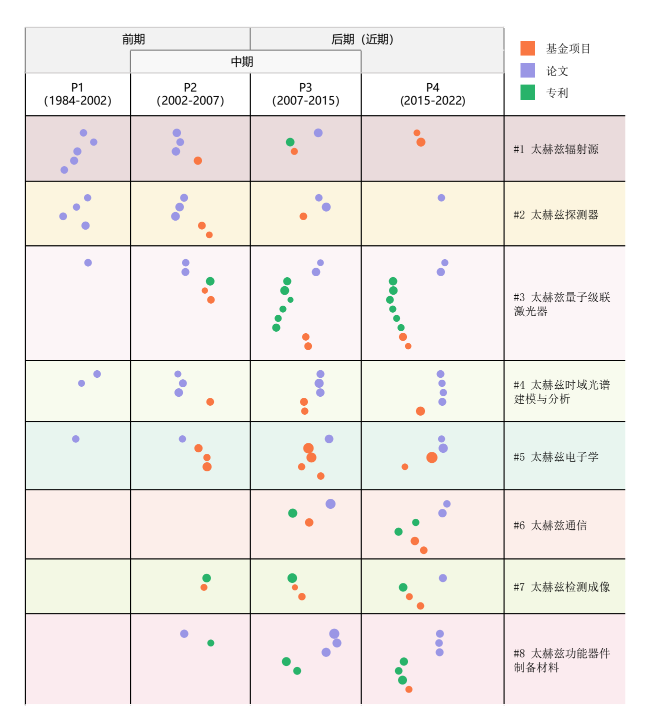


图 10 领域主题社区节点数量时序分布

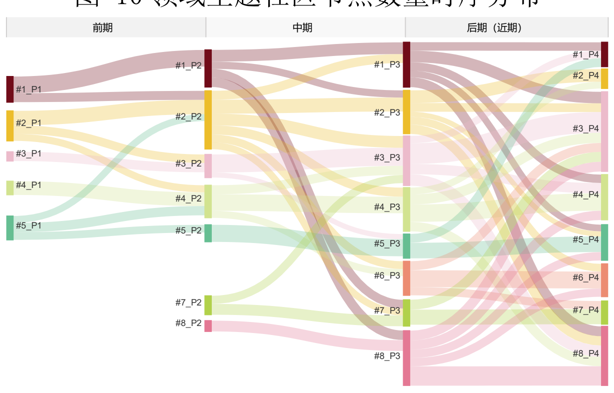


图 11 领域主题社区时序关联图

从社区内外主题演化关系（图 11）来看，总体上社区间研究分裂和融合的趋势显著增强。#1 太赫兹辐射源和#2 太赫兹探测器难以维持研究社区的稳定，主题分裂产生的外溢效应明显，为其他多个主题社区的发展提供基础。#3 太赫兹量子级联激光器前期主要融合#2 太赫兹探测器的部分内容，中期广泛融合多社区知识，后期分裂和融合趋势愈发显著。#4 太赫兹时域光谱建模与分析前期和后期分裂和融合现象均较为明显，后期分裂和融合的广度有了较大提升。#5 太赫兹电子学前期的分裂现象相对明显，中期与其他社区联系较为松散，后期则融合了#1 太赫兹辐射源、#2 太赫兹探测器、#8 太赫兹功能器件制备材料的多社区内容，知识范畴得到进一步扩展。#6 太赫兹通信、#7 太赫兹检测成像和#8 太赫兹功能器件制备材料社区形成较晚，对其他社区研究内容的融合能力强，尤其#6 太赫兹通信的产生直接受益于#2 和#4 的分裂扩散；同时在领域发展的后期#8 的分裂和融合现象均得到显著增强，知识扩散和聚合能力同步提升。

（2）演化内容特征分析

利用所有社区全部时间段的主题词构建共词网络，将各社区不同时间段的主题节点聚合成不同的主题簇，根据 3.3.4 所述方法计算不同主题簇的密度和向心

度，以中位数为分界线，构建战略坐标图（图 12）。

第一象限中仅包括第四阶段的#5 太赫兹电子学，其他主题簇集中分布在第三和第四象限。#1 太赫兹辐射源和#7 太赫兹检测成像在发展的所有阶段均位于第三象限，为边缘非稳定类主题簇，但#7 向心度逐渐增长，自身发展对领域影响逐渐扩大，而#1 向心度呈逐渐降低趋势，日趋走向边缘地带。#2 太赫兹探测器前期密度和向心度同步提升，中后期两个指标均持续降低，自身稳定性以及对其他社区的影响力逐渐减小。#3 太赫兹量子级联激光器、#4 太赫兹时域光谱建模与分析、#5 太赫兹电子学、#6 太赫兹通信密度和向心度呈现同步增长趋势，逐渐走向成熟，在领域发展中进入愈发核心的位置，其中#3 太赫兹量子级联激光器在领域发展后期阶段向心度最高，成为最核心的主题簇；#5 的密度变化尤其显著，在第四阶段成为密度最高主题簇，研究内容逐渐进入成熟期。#8 太赫兹功能器件制备材料在发展的过程中，密度减小、向心度增加，表明其在发展的过程中逐渐不稳定，研究内容变化较大，同时与其他研究分支联系加深，逐渐步入核心位置。



图 12 领域各主题簇战略坐标图

综合主题社区演化形式特征和内容特征分析发现：

#1 太赫兹辐射源和#2 太赫兹探测器社区起步早，研究内容逐渐收缩，近年来分裂现象愈发显著，不稳定性加剧。值得注意的是，#1 在近期均为基金项目主题，其未来发展值得继续关注。

#3 太赫兹量子级联激光器起源较早，随着领域的发展快速扩张，近期分裂和融合趋势显著，处于领域发展的核心位置，在专利文献中研究基础扎实。

#4 太赫兹时域光谱建模与分析和#5 太赫兹电子学研究探索较早，处于平稳发展阶段。其中#4 近期分裂和融合的广度均位于前列，核心性和稳定性得到持

续提升，论文是其主要发展阵地；#5 近期对其他社区的知识融合趋势明显，自身发展走向成熟期，基金项目在发展中后期起到关键作用。

#6 太赫兹通信启动最晚，正在快速扩张，对其他社区研究内容的融合能力强，近期分裂现象明显，影响多个社区的发展，已成为领域相对核心地位的主题簇，近期同时受到基金项目、论文和专利的广泛关注。

#7 太赫兹检测成像开始较晚，处于持续扩张期，分裂和融合现象增强，自身稳定性不足，核心性发展停滞，近期主题节点中基金项目、论文和专利均有涉猎，需要进一步关注其成长状况。

#8 太赫兹功能器件制备材料研究开启较晚，持续快速扩张，分裂和融合现象愈发显著，自身稳定性有所降低，核心性不断增强，近期发展受到基金项目关注，论文和专利的研究成果稳步发展，未来发展前景广阔。

5 结语

本文提出基于多源数据的领域主题演化研究框架，并在美国太赫兹领域展开实证研究。引入基金项目、论文和专利数据，丰富数据来源；利用多维有序样本聚类方法，对不同数据源分别进行时间窗口划分，构建科学合理的数据集；提出改进的词袋构建方法，提高主题识别结果的可读性和可理解性；利用 Louvain 社区发现算法，融合多源数据形成的研究主题，聚合成关系紧密的主题社区，简化主题演化网络，清晰展现主题演化路径；从演化形式特征和内容特征两个角度综合分析主题社区演化趋势。为验证研究框架的有效性和实证研究结果的可靠性，本论文咨询了太赫兹领域的专家，专家认为所提出的数据集划分方法与领域实际发展阶段吻合：第一阶段（1984-2002 年）为领域发展起步阶段，研究项目大多为基础科学研究类型，数量和经费均比较少；第二阶段（2002-2007 年）为初步发展阶段， 2004 年美国将太赫兹科技评为“改变未来世界的十大技术”之一；第三阶段（2007-2015 年）为快速发展阶段，该阶段处于领域发展的黄金时期，领域内部各个技术方向百花齐放，研究方向较为发散；第四阶段（2015-2022 年）为发展成熟阶段，科研投入更加注重太赫兹技术的应用价值，研究项目数量下降的同时单个项目的经费大大提升。同时，专家认为文本挖掘的主题社区基本涵盖了太赫兹研究的各个领域，主题演化分析直观、有效地揭示太赫兹领域发展态势及各阶段发展特征。因此，基于本文提出的研究框架，可以一定程度上解决当前研究中存在的数据来源缺乏、多源数据融合困难、数据划分客观性不足、主题挖掘结果可读性和可理解性差、领域主题演化脉络不清晰等问题。

本研究在数据源的全面性和方法的适用性存在一些局限性，可进一步开展以下几个方面的研究：①本文重点关注三种数据源，后续可扩展更丰富的数据来源，进行更全面的领域主题演化分析；②本文仅在美国太赫兹领域开展实证研究，后续可在其他研究领域进行实证分析，验证研究方法的适用性。

参考文献

[1] 许海云，董坤，隗玲，等.科学计量中多源数据融合方法研究述评[J].情报学报,2018，37（03）：318-328.

[2] 李广建，杨林.大数据视角下的情报研究与情报研究技术[J].图书与情报,2012（06）：1-8.

[3] XU H Y, YUE Z H, WANG C, et al.Multi-source data fusion study in scientometrics[J].Scientometrics,2017，111（2）：773-792.

[4] 谭晓，李辉.基于多源数据知识融合方法的研究前沿识别[J].现代情报,2019，39（08）：29-36.

[5] 冯佳，穆晓敏，王伟.面向研究前沿识别的载体-特征-关系融合模型研究[J].图书馆杂志,2020，39（09）：56-63.
[6] WANG X.Research on the discourse power evaluation of academic journals from the perspective of multiple fusion: Taking Medicine, General and Internal journals as an example[J].Journal of information science，0（0）：01655515221107334.
[7] 陈启明，王效岳，白如江，等.多源数据融合下突发公共事件社会关注与政策趋向研究——以新冠肺炎疫情为例[J].情报探索,2022（06）：15-25.
[8] 胡吉霞. 面向多源数据的学科知识网络构建方法与应用研究[D]. 西安电子科技大学,2021.
[9] 王春秀，冉美丽.学科主题演化定量分析的理论基础探析[J].现代情报,2008（06）：48-50.
[10] 梁爽，刘小平.基于文本挖掘的科技文献主题演化研究进展[J].图书情报工作,2022，66（13）：138-149.
[11] 陈悦，刘则渊，陈劲，等.科学知识图谱的发展历程[J].科学学研究,2008（03）：449-460.
[12] MORRIS S A，YEN G，WU Z，et al.Time line visualization of research fronts[J].Journal of the American Society for Information Science and Technology,2003，54（5）：413-422.
[13] PALLA G，BARABASI A L，VICSEK T.Quantifying social group evolution[J].Nature,2007，446（7136）：664-667.
[14] 周源，张超，唐杰,等.基于主题变迁的领域发展路径智能化识别——以人工智能为例[J].图书情报工作,2018，62（14）：62-71.
[15] 陈悦，王康，宋超，等.一种用于技术融合与演化路径探测的新方法：技术群相似度时序分析法[J].情报学报,2021，40（06）：565-574.
[16] 刘怀兰，刘盛，周源，等.基于多源文本挖掘的技术演化路径识别[J].情报理论与实践,2022，45（11）：178-187.
[17] MEYER M.Tracing knowledge flows in innovation systems[J].Scientometrics,2002，54（2）：193-212.
[18] 刘自强，许海云，岳丽欣，等.面向研究前沿预测的主题扩散演化滞后效应研究[J].情报学报,2018，37（10）：979-988.
[19] 李慧，孟玮.专利视角下的美国空军核心技术演化分析[J].情报理论与实践,2021，44（02）：41-49.
[20] 李慧，胡吉霞，佟志颖.面向多源数据的学科主题挖掘与演化分析[J].数据分析与知识发现,2022，6（07）：44-55.
[21] FISHER W D.On grouping for maximum homogeneity[J].Journal of the American Statistical Association,1958，53（284）：789-798.
[22] 李俊，毕华兴，李笑吟，等.有序聚类法在土壤水分垂直分层中的应用[J].北京林业大学学报,2007（01）：98-101.
[23] 大布穷，叶彦辉，赵垦田.西藏色季拉山急尖长苞冷杉生长规律研究[J].安徽农业科学,2010，38（17）：9317-9320+9344.
[24] 张多，韩逢庆.基于支持向量机和有序聚类的岩层识别[J].智能系统学报,2014，9（01）：98-103.
[25] 祖坤琳，赵铭伟，林鸿飞.基于有序聚类的专利知识演化研究[J].计算机工程与科学,2016，38（04）：785-791.
[26] 严广松，路允芳.多维有序样本的聚类方法研究[J].统计与决策,2008（04）：29-30.
[27] DU Y J，YI Y T，LI X Y，et al.Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet allocation[J].Engineering applications of artificial intelligence,2020，87：13.
[28] 谭春辉，熊梦媛.基于 LDA 模型的国内外数据挖掘研究热点主题演化对比分析[J].情报科学,2021，39（04）：174-185.
[29] 张学成，周斌，孔瑞远，等.大型仪器利用情况调查数据异常值检测的数学方法比较[J].数学的实践与认识,2012，42（11）：50-54+56-57+55.
[30] 刘路. 基于 Louvain 算法的社区发现与核心节点挖掘研究[D]. 西安电子科技大学,2021.

[31] NEWMAN M E J.Modularity and community structure in networks[J].Proceedings of the National Academy of Sciences of the United States of America,2006, 103 (23): 8577-8582.
[32] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al.Fast unfolding of communities in large networks[J].Journal of statistical mechanics-theory and experiment,2008: 12.
[33] 隗玲，许海云，胡正银，等.学科主题演化路径的多模式识别与预测——一个情报学学科主题演化案例[J].图书情报工作,2016, 60 (13): 71-81.
[34] 唐果媛.基于共词分析法的学科主题演化研究方法的构建[J].图书情报工作,2017, 61 (23): 100-107.
[35] 周毅. 模因视角下知识网络的主题演化研究[D]. 兰州交通大学,2021.
[36] 姜鑫，王德庄，马海群. 社会网络分析方法在图书情报学科的应用研究[M]. 北京：知识产权出版社，2019.
[37] LEE B, JEONG Y I.Mapping Korea's national R&D domain of robot technology by using the co-word analysis[J].Scientometrics,2008, 77 (1): 3-19.
[38] SATOPAA V, ALBRECHT J, IRWIN D, et al. Finding a kneedle in a haystack: detecting knee points in system behavior[C]// International conference on distributed computing systems workshops. IEEE Computer Society, 2011.

(通讯作者：朱相丽　E-mail: zhuxl@mail.las.ac.cn)

作者贡献说明：

张敬：论文构思、数据管理和分析、初稿写作；

朱相丽：论文框架调整完善，论文的指导、审核与修改。

感谢北京邮电大学亓丽梅教授和中国科学院空天信息创新研究院李超研究员对本文提供的专业意见。